LUND UNIVERSITY
School of Economics and Management

Is there macroeconomic predictive power in Swedish business news?

An exploratory analysis of how business newspaper articles can be used for prediction of major Swedish macroeconomic variables.

by

Erik Ris ine15eri@student.lu.se

NEKH01
Bachelor's Thesis (15 ECTS)
(August) (2021)
Supervisor: Joakim Westerlund

Abstract

This thesis explores if, and how, what is written in the newspaper can be used to forecast macroeconomic variables such as inflation, unemployment and consumption. A large data set consisting of articles from the largest Swedish business newspaper is transformed using different methods from the Natural Language Processing field. The focus lies on topic modelling by Latent Dirichlet Allocation as well as sentiment analysis. The newspaper data are represented as a combination of the distribution over topics covered in the newspaper as well as sentiment scores of prominent articles. The data representations of the newspaper are created, explored and later used to predict the movement of the economic variables using Lasso regression, a method automatically selecting important input variables. The newspaper data on stand-alone basis have not been shown to have predictive power for these macroeconomic variables. But, when allowing the model to also be trained on the lagged economic variable interesting observations are made. The predictive performance is improved by the newspaper data, in comparison to only using the lagged economic variable. This is the case for *expected inflation*, *CIPF-inflation* and *unemployment*.

This bachelor's thesis was written under guidance of Lund University School of Economics and Management, Department of Economics - Econometrics and in collaboration with the company Sanctify Financial Technologies.

Keywords: Topic modelling, LDA, Sentiment Analysis, Swedish Newspaper Data, NLP, Economics, Lasso Regression

Acknowledgements

I would like to express my gratitude to my supervisor Joakim Westerlund, professor at Lund Univeristy School of Economics and Management, Department of Economics who have guided me in the process, and answered my questions even on the sunniest days of his summer vacation.

This thesis is written in collaboration with Sanctify Financial Technologies, and I would like to thank both Oscar Dahlblom and Gustav Johnsson Henningsson, my two primary supervisors for their interest and help. Also, a special thank you to my former master thesis partner Axel Sjöberg for the many interesting discussions during our collaborative work, and for helping laying the foundation to this thesis.

A final thank you to my dear friend Oskar Stigland for always being the sparring partner I need when tackling difficulties in Python, and for introducing me to machine learning.

Contents

Introduction

Natural language processing (NLP) is a growing field of research with successful applications in a vast number of domains. NLP aims to make textual data interpretable for computers. In order to make sense of a large body of text, a common approach is to use a topic modelling algorithm. Similarly to how humans decompose a document, the method is designed to find the topics covered in a text. This yield an interpretable text data representation, where a large text corpus is mapped to a specified number of topics. In this work the topic model of choice is Latent Dirichlet Allocation (LDA), which is a leading method for topic modelling used for text analysis in many research fields. In economics, however, the applications have been sparse according to Thorsrud (2018).

This thesis is an extension to a recent master thesis in which Swedish business newspaper data is used to predict the movement of Swedish stock market indices. The results obtained where cautiously optimistic, but more research required, for further explanation, see Ris and Sjöberg (2021). In this work a similar approach will be taken including some critical alternations, with the main differences being the response variables and their update frequency, the methods used to construct the time series representation and the type of predictions made.

To what extent can topics and sentiments obtained from the written newspaper be used to forecast economical variables? Can data representations created by LDA and sentiment analysis in combination help forecasting either: expected inflation, actual inflation, business confidence, consumer confidence, consumption or unemployment?

The purpose of this thesis is to investigate the predictive power of Swedish newspaper data in relation to major variables connected to the Swedish economy. A large text data set is used to form input signals based on the themes covered in the newspaper. Topics extracted by LDA are turned into time series and later scaled and sign adjusted using the sentiment scores from the most important articles. This, to test the usefulness of Swedish newspaper data for forecasting the following macroeconomic variables: Expected inflation, actual inflation (CPI and CPIF), a business confidence index (BCI), a consumer confidence index (CCI), consumption and unemployment. Including the lagged version of the macroeconomic variable in the predictive model yield positive results for several of the economic variables, indicating the usefulness of newspaper data for making macroeconomic predictions.

In *Appendix A* and *Appendix B* the theory used in this work is described and explained, specifically separated into *Machine Learning* and *Natural Language Processing*. In *Chapter 1* previous research from relevant domains are covered, focusing on the intersection between NLP and economics. In *Chapter 2* the data used are described and analyzed. In *Chapter 3* the research methodology of this work is presented thoroughly in order to increase understanding and simplify replication. The results of the predictive analyses are also presented and discussed. Finally, the conclusions drawn from the results are presented alongside suggestions for future research in *Chapter 4*.

1

Previous Research

1.1 NLP - LDA in Economics

In a series of research papers and publications Leif Anders Thorsrud and Vegard Høghaug Larssen have explored the application of newspaper data in economics. Their main tool for examining the newspaper data and extracting input features is topic modelling by LDA. Using LDA, a large text data set is transformed into time series telling the story of what is covered in the newspaper. The time series are the topic distributions obtained from the LDA algortihm on a daily basis. The publications operates under the hypothesis that the more intensive a given topic is represented in the newspaper at a given point in time, the more likely it is that this topic represents something of importance for the current state of the economy. The topic time series are often also complemented with a sentiment score extracted from the most prominent article for each topic each day. The data set used in most of their publications originates from the Norwegian newspaper Dagens Næringsliv, which is the largest business newspaper in Norway. See Larsen and Thorsrud (2019), Thorsrud (2018), Larsen and Thorsrud (2017) and Larsen et al. (2020) for examples. Some of these publications are further covered below.

In Thorsrud (2018) the newspaper data set is used in order to construct a latent threshold model used to nowcast Norwegian GDP and to construct a news-based business cycle index.

Larsen and Thorsrud (2019) investigate how a Norwegian newspaper corpus consisting of several years of data for the largest business newspaper in Norway is connected to macro indicators for the Norwegian economy. Output, investment, consumption, total factor productivity, asset prices and business confidence are the variables examined. The news corpus is decomposed into topic time series. Each macro variable time series is modelled as an AR(p) (auto regressive model of order p) and compared to regressor augmented AR(p) models, one for each topic time series in combination with the variable of interest. The time series are compared using statistical tests in order to examine if the given topic time series add to the forecasting capability of the model compared to using the AR(p) model of the economic variable. This

modelling is conducted both pseudo in-sample, where the topic model is trained on all text data and out-of-sample, where the topic model is only trained on the training partition of the data set. The results are promising, showing relationships between all of the variables to at least some topics. Output and consumption are shown to be related to most topics. This initial modelling is conducted to be able to learn which topics to include when constructing a news-based business cycle index.

In Larsen et al. (2020) the relation between American news and inflation expectations as well as actual inflation is examined. Approximately five million news articles stretching from 1990 to 2016 are used to construct 80 time series covering to what extent different topics have been written about in the newspapers. The news data set originates from the Dow Jones News Archive and consists of many different news outlets, including the Wall Street Journal. The number of topics $K = 80$ is consistent with earlier works from the authors. To form the time series LDA is performed, and a simple dictionary based sentiment analysis is conducted to find the sign of the distribution weight for each topic and day. The inflation expectations and actual inflation are modelled and predicted using Lasso regression, to interpret the results R^2 as a measure of explained variability is used.

1.2 NLP - Swedish Newspaper Data

In Blad and Svensson (2020) topic modelling methods of Swedish newspaper data are explored and compared, both qualitatively and quantitatively. LDA and Non-negative Matrix Factorization (NMF) are used to construct topics for a large data set of Swedish text data from the publishing company *Bonnier News*, and the interpretability of the topics are compared. The data set includes data from Dagens Industri, but also other large newspapers and magazines. It is shown that lemmatization is preferred over stemming for Swedish (as well as other Germanic languages). It is also shown how experts in the field prefer topics constructed using only nouns or nouns and proper nouns over topics constructed using all word classes, and how LDA using Gibbs sampling is preferred over LDA using Online Variational Bayes as inference technique.

In (Ris and Sjöberg, 2021) the the newspaper data set from Dagens Industri used in this thesis is examined and explored in relation to market predictions. The data set is used to construct interpretable input data representations used for stock market index predictions. The representations are created by a combination of *Latent Dirichlet Allocation* and *Sentiment Analysis*. It is shown that full length articles seems more suitable to construct representations for long term predictions. Similarly to for example Larsen and Thorsrud (2019) and Thorsrud (2018) $K = 80$ is preferred over finding less topics in the corpus. It is also shown that it is more effective to omit some of the less "informative" topics subjectively before conducting the following feature selection and predictive modelling. This is a similar approach taken in Thorsrud (2016), where subjective selection of topics is undertaken in order to keep only informative topics when modelling.

2

Data

2.1 Newspaper Data Set

The newspaper data used in this thesis originates from the daily printed version of Dagens Industri, the largest business newspaper in Sweden, (Dagens Industri, 2020). Similarly to Larsen and Thorsrud (2019), one data source is deemed enough. This as Sweden and Norway are rather small and similar economies with only a handful business newspapers. The largest among these is believed to cover most of the business information published at each point in time. Using only one data source is also a more simple approach compared to collecting and processing all published business news, as the data set already is large and computationally demanding to handle. The data set consists of around 180,000 news articles published during the time period 2007 - 2016. This results in 2,900 days or 120 months of newspaper data.

The training data set is constructed of the first eight years of the data set: 2007-2014 and the test data set is constructed of the two final years: 2015 and 2016. This result in a 80 % - 20 % split of the data set, a common ratio used when partitioning a data set into training and testing. From the training data partition a validation data set is constructed, used for pseudo out of sample testing for the input data selection.

The data set is downloaded as text files of 500 articles per file from Retriever Mediaarkivet, (Retriever, 2021). These text files are read into a Pandas data frame. In this process a rough first data cleaning is carried out, URL:s, symbols (except for punctuation) and numbers are removed. Resulting in a final data set where each article is represented only by letters and punctuation. This data set is stored and easily processed as a Pandas data frame.

2.2 Exploratory Data Analysis

In Ris and Sjöberg (2021) a thorough exploratory data analysis of the newspaper data set is presented. The data set originally includes articles published during 2007-2020, however an anomaly for 2017, 2018, 2019 and 2020 is detected. This anomaly refers to the preambles of the newspaper articles being compromised in the data base. Also, an anomaly regarding some articles being duplicated in the data set is detected and curated. In Ris and Sjöberg (2021) different data sets are compared and the shorter length data set (2007-2016), not including the years with compromised preambles, outperforms the full length data set (2007-2020). This discrepancy in performance could be due to the anomaly of the preambles for the most recent years, (Ris and Sjöberg, 2021). To remove this potential bias in the data, the text data of choice in this work is therefore the shorter length data set with coherent and uniform articles.

2.3 Economical Variables

Variable	Description	Data Source
Expected Inflation	Inflation expectations in 12 months	KI
BCI	Business Confidence Index	OECD
CCI	Consumer Confidence Index	OECD
Consumption	Consumption compared to 12 moths ago	Ekonomifakta
Unemployment	Unemployment rate seasonally adjusted	FRED
Inflation	Current inflation (CPI and CIPF)	Ekonomifakta

Table 2.1: Shorthand, description and source presented for the the target variables.

In Table 2.1 the response variables examined in this work are summarized. The choice of target variables stem from Larsen et al. (2020) and Larsen and Thorsrud (2019) as well as discussions with Sanctify Financial Technologies and data availability. The latter reason as most prominent macro economical variables are available only with a quarterly update frequency.

The expected inflation data are aggregated and made available by Konjunkturinstitutet (KI), and stem from surveys conducted. The data set is downloaded from Konjunkturinstitutet (2021). The current inflation data and consumption data (change in consumption, trend adjusted) are constructed and publicized by Statistics Sweden (SCB) and made available by Ekonomifakta, a site organized by Svenskt Näringsliv. Ekonomifakta summarize and compile Swedish financial and economical data from SCB, but also other well known sources such as Organisation for Economic Co-operation and Development (OECD) and EuroStat, (Svenskt Näringsliv, 2015). The unemployment data originates from OECD and made available by the Federal Reserve Economic Data (FRED) database and downloaded from OECD (2021c). The BCI data and CCI data are constructed by and made available by (OECD) and downloaded from OECD (2021a) and OECD (2021b). All economic variable data

sets are downloaded and variables of interest are extracted and stored in a Pandas data frame.

In Figure 2.1 the response variables are plotted for the training period. The plots are showing some differences between the data sets, for example are the BCI, CCI and consumption time series smoother compared to the other data sets.

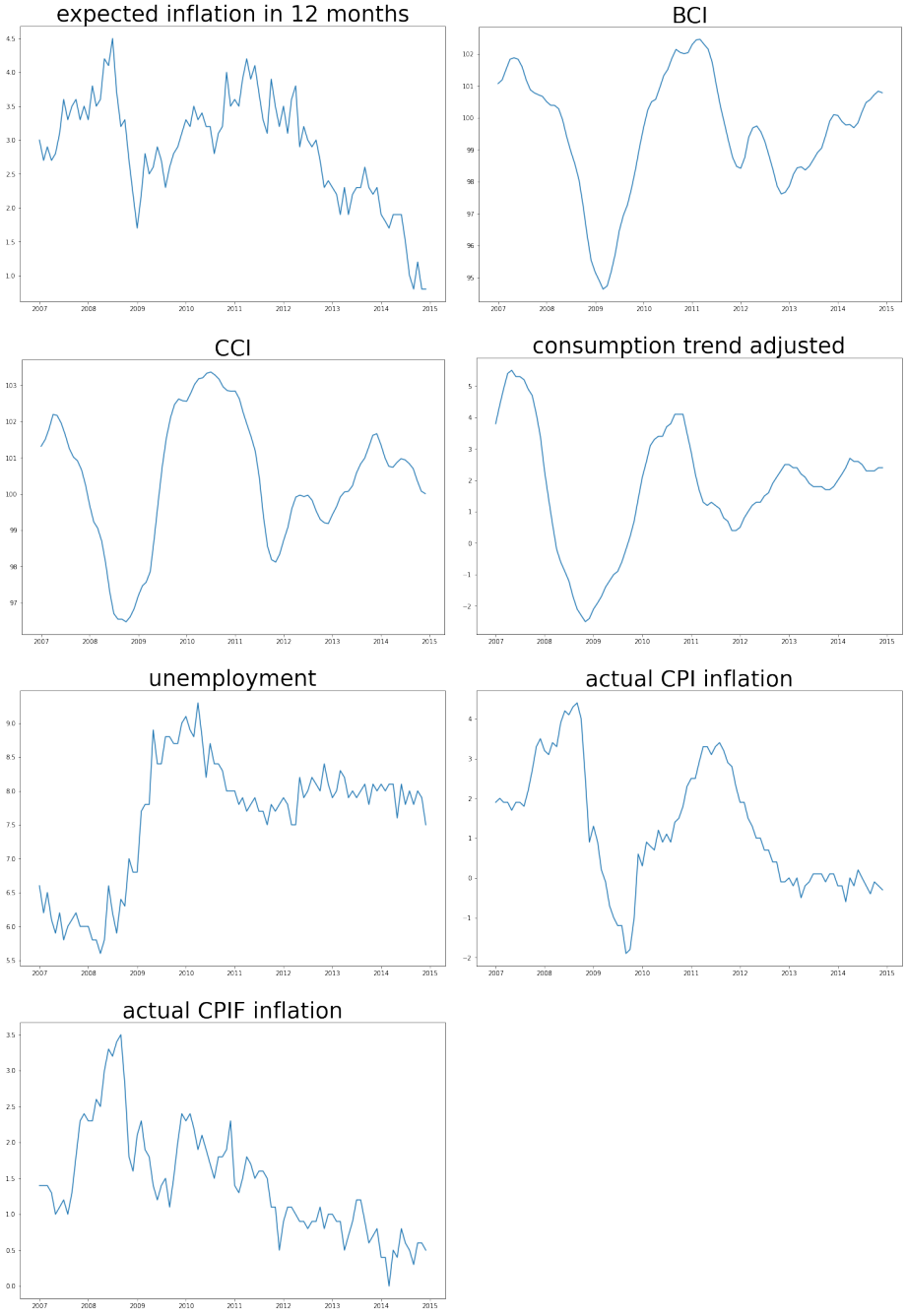

Figure 2.1: Expected inflation, BCI, CCI, consumption, unemployment and actual inflation (CPI and CPIF) plotted in order for the training period.

3

Empirical Analysis

3.1 Data Processing

The newspaper article data set is first cleaned as described in Chapter 2. This first rough cleaning of the text data yields a data set where the articles are made up of sentences without any special characters, numbers or URL-links. Keeping the punctuation in the articles at this point enables the construction of *n-grams* without overlapping sentences in a later stage of the process. The next step in the process is to remove stop-words from the articles, this is done by omitting words in a predefined list of Swedish stop-words. After the removal of the stop-words the process continues with tokenization of all of the words in each article. This is conducted by using a pre-trained tokenizer from UDpipe using the spcay-udpipe package, (Take Lab, 2020). In the tokenization step the part of speech (POS) or word class is extracted for each word and later used to lemmatize the word and for filtering out nouns. Lemmatization is conducted in order to map all versions of a word regardless of tense, conjugation or genitive to the lemma, or basic form of the word. The articles are pre-processed differently in order to construct one data set prepared for the topic modelling and one data set for the sentiment analysis. This pre-processing step and data partition is conducted in line with, Ris and Sjöberg (2021).

3.2 Pre-processing for Topic Modelling and Sentiment Analysis

The first step in order to prepare the data set for the topic modelling is to construct and add *n-grams*, in this work n is set to two, similarly to Ris and Sjöberg (2021) resulting in *bigrams*. These are created on a sentence level, not allowing for *bigrams* with words from different sentences. The bigrams are created by attaching two following words with an underscore to create one token.

Blad and Svensson (2020) show that nouns or nouns in combination with proper nouns are preferred over using all word classes when creating topics. Based on this finding all words not being nouns and all bigrams not containing at least one noun are omitted from the topic modelling data set.

The tokens in the data set are then lemmatized using a pre-trained lemmatizer for Swedish texts from the lemmy package, (Kristiansen, 2019). The final processing step of the texts is to remove frequent and infrequent tokens. This is done to reduce the size of the text data set, but also since words occurring very frequent or infrequent does not contribute specific information. Removing tokens occurring more than 50,000 times in the corpus and tokens occurring less than 5 times similarly to Ris and Sjöberg (2021) reduces the size of the vocabulary by around 90 %. This procedure is carried out in order to reduce the noise in the data set only keeping words adding information, as well as simplifying the training of the topic model. Examples of tokens occurring more than 50,000 times, thus being omitted, are tokens referring to Dagens Industri and including "di". These tokens are assumed not to add any information beyond the origin of the news.

The last step before the text data is ready for the topic modelling is to create a bag of words representation of the text corpus, which is the required input representation for the LDA-algorithm used in this work.

The financially focused sentiment lexicon constructed in Loughran and Mcdonald (2011) is translated to Swedish, tokenized and lemmatized. The words in the tokenized article data set are lemmatized in order to match the representation of the sentiment lexicon and therefore increase the "hit rate". The lemmatization is again carried out using the pre-trained lemmatizer from the lemmy package (Kristiansen, 2019).

3.3 Feature Construction

The topic modelling in this work is conducted using LDA. The implementation of the algorithm used is created by Mallet, McCallum (2002). It is an implementation of the LDA algorithm using Gibbs sampling as inference method, and chosen over an implementation using Online Variational Bayes based on the findings of Blad and Svensson (2020). To access the Mallet implementation, originally written in *Java*, a wrapper from the *Gensim* package is used. The topic modelling is conducted using mostly default parameters following (Blad and Svensson, 2020), the number of topics and iterations of the Gibbs sampling are set manually. The number of iterations is increased from 1,000 as default to 2,500 based on suggested settings from (McCallum, 2002). The number of topics is set to 80 based on findings in Larsen and Thorsrud (2019) and Thorsrud (2018) as well as in Ris and Sjöberg (2021).

The pre-processed topic model data set is fed to the LDA algorithm and the output is a trained topic model, which is used to construct a topic distribution on article basis, henceforth called the *article distribution* data set. Each topic's word distribution is

also extracted, and used to create topic representations using the most important words for each topic. In line with Ris and Sjöberg (2021) and Larsen and Thorsrud (2019) the full text data set is used as input data in the LDA model. This can result in a look ahead bias in the topics created by the LDA model, if these are used for out of sample prediction, but deemed necessary as a starting point for examining the usefulness of the topic data representation in this work. Continuous updating of the topic model is also computationally very costly, and not a suitable approach in this thesis.

Figure 3.1: Topic 60 - topic 79 presented as Wordclouds.

In figure 3.1 20 random topics obtained by training the topic model on the text data set on article basis are presented, the topics can be said to be of varying quality from a human reader point of view. The topics are presented in Swedish and by using the ten most important words from each topic's word distribution. Topic 75 for example seems to only concern weeks and topic 61 "last year", while topic 78 concern the state of the economy and topic 64 seems to concern American politics.

The sentiment analysis is carried out using a lexical approach in order to find the tone of each newspaper article. The word list used in this work is the financial specific sentiment lexicon constructed in Loughran and Mcdonald (2011). The sentiment score of article i is constructed as the crude measure described in (3.1) where $W_{positive}$ and $W_{negative}$ are the words in the article labeled positive and negative by the sentiment dictionary and W_{total} are the total number of words in the article. The scaling of the sentiment score by the total word count of each article is conducted in line with Thorsrud (2016).

$$Sentiment_i = \frac{W_{positive} - W_{negative}}{W_{total}} \tag{3.1}$$

3.4 Time Series Representation

Based on the creation of a monthly time series representation of the topics extracted from a newspaper data set in Larsen et al. (2020), the output from the topic modelling and sentiment analysis are combined. By combining all articles each day to a new document, the topic model trained on the data set on article basis is used to extract daily topic distributions. That is, "to what extent is topic k covered in the newspaper today?". This results in a data set of 80 topic time series on a daily basis where the topic weights each day by construction sum to one, from here on called the *daily distribution* data set.

Using the topic distribution from the article distribution data set, the most important article for each topic each day is extracted. The sentiment score for this article is then used to sign adjust and sentiment scale the topic time series in the daily distribution data set, this again based on Larsen et al. (2020). The scaled sentiment score for the most important article for each topic each day is multiplied with the topic weight in the daily distribution data set. To transform this data set from daily frequency to monthly a simple aggregation of the topic sentiment scores per day are conducted by summing all scores each month, finally resulting in the final monthly distribution data set.

Several other time series representations are created, and later compared in the predictive validation phase. The four time series representations shown to be successful in the validation phase, and thus, named can be seen in Table 3.1.

- The "second" approach presented in Ris and Sjöberg (2021) is created using the scaled sentiments and the daily distribution data set. This approach is shown to be the most successful when different representations are compared in Ris and Sjöberg (2021). A variation to this approach using the sentiments and topic distributions from the top five most important articles per topic is created as well. These approaches only account for the most important article, and the top five most important articles each day respectively.
- A variation to the "first" approach presented in Ris and Sjöberg (2021) is created. This approach was originally based on multiplying all topic weights for each article with the article sentiment score and then summing all topic-sentiment scores each day, this however, was not especially successful. In the alternation the topic distribution weights in the article distribution data set are squared before being multiplied with their respective weighted sentiment score and summed for each day. This in order to again only account for the most important topics per article, by squaring the distribution weights the smaller values is further reduced while the larger values are less reduced. The

monthly variant of this data set is henceforth called *squared sum* distribution data set, as the summation approach now also is squared.

– Variations to the creation of the *daily distribution* data set using a moving average filter is created. When using a monthly moving average filter before summing all days of each month, this data set is called the *moving average* distribution data set.

– A different daily representation data set is created by using the data set where all articles each day are combined to one document. This data set, consisting of a one document containing all publicized articles each day, are then fed to the topic model. The new trained topic model is used for inferring topics on this data set, already on daily basis. The settings for the LDA algorithm are not altered, and this method results in a different set of topics. This data set of daily topic distributions is not sign adjusted using the sentiment scores. The monthly input data set is hence created as the summation of the topic weights each month, and is henceforth called *daily document* data set.

Name	Description
Monthly Dist.	Data representation created following (Larsen et al., 2020)
Squared Sum	Following (Ris and Sjöberg, 2021), but squaring topic dist.
Moving Avg.	Similar to *Monthly Dist.* but including a monthly MA-filter
Daily Doc.	Similar to *Monthly Dist.* but documents per day in topic model

Table 3.1: Overview of the four different topic-sentiment time series representations proven successful in the validation phase.

In order to explore a sample of the topic-sentiment series constructed as described in the first and second paragraph of Section 3.4, three topics concerning the economy are plotted below in different formats. The three topics plotted are described by their seven most important words as shown in Table 3.2, the full ten words can be seen is the appendix. In Figure 3.2 three topic time series sign adjusted by the scaled sentiment score are presented on a daily basis. In Figure 3.3 the same three topics are again presented but now in the monthly representation from the monthly distribution data set. In Figure 3.4 the topics are again plotted, now using a 150 days moving average filter to smooth the time series and better capture large movements.

Topic	Important words
46	bankruptcy, source , loan, information, owner, in_bankruptcy, financing
71	bank, loan, big_bank, to_bank, bank_and, lending, capital
78	economy, growth, forecast, recovery, next_year, economic_situation, sign

Table 3.2: Topics 17, 46 and 78 described by their seven most important words, translated from Swedish to English.

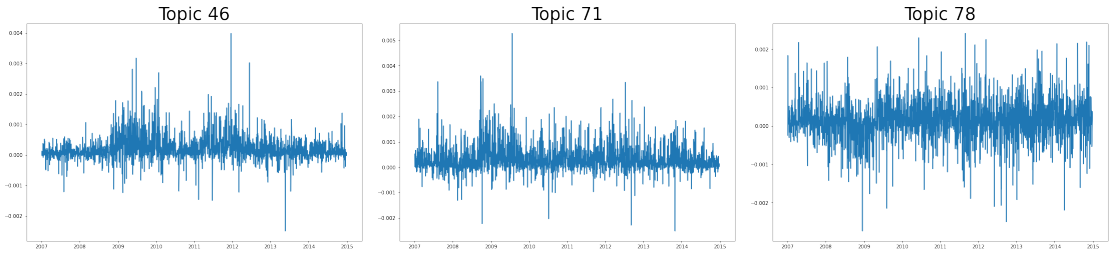

Figure 3.2: Topic 46, topic 71 and topic 78 plotted in order for the training period on a daily basis.

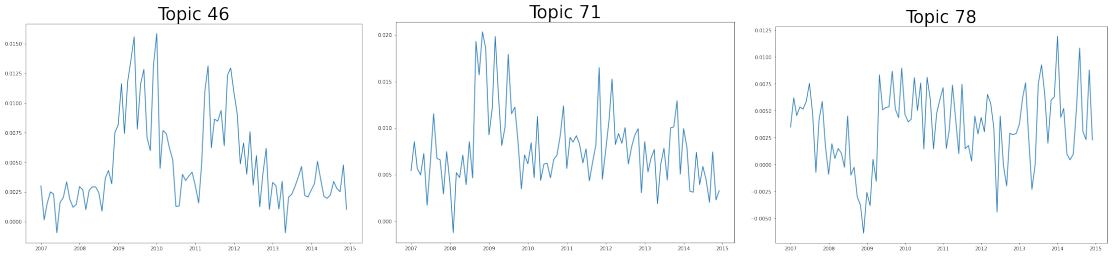

Figure 3.3: Topic 46, topic 71 and topic 78 plotted in order for the training period on a monthly basis.

Figure 3.4: Topic 46, topic 71 and topic 78 plotted in order for the training period using a 150 days, approximately 6 months, moving average filter.

As can be seen in Figure 3.3, and even better in Figure 3.4, Topic 46 concerning bankruptcies and loans spike both during the 2007-2009 global financial crisis, as well as later during what could be during the financial crisis in Spain, Portugal and Greece. Topic 71, concerning big banks and loans also spike during the 2007-2009 financial crisis. Topic 78, concerning the economic situation has a clear negative drop starting in the middle of 2007 and stabilize again in the beginning of 2009. In the more recent years however, this topic sentiment score is more volatile. These characteristics are also distinguishable in Figure 3.2, even if not as prominent. This sub sample of the topic sentiment time series seem to make sense as important events in the economy are captured and clearly distinguishable.

Based on findings in Thorsrud (2018) adjusting the topic-sentiment time series for a linear trend does not affect the final result. Thus, any such adjustment is omitted in this work. And, based on findings in Ris and Sjöberg (2021) as well as examining the covariance matrix of the sentiment scaled topic time series there seems to be no apparent co-linearity issues for the input data set.

3.5 Feature Selection and Processing

The predictive analysis in this work is conducted using Lasso regression in order to automatically select important features. The process is divided into a training and validation phase used to decide which input data representation to use, and a test phase where the selected input data representations are used to predict on the unseen test data. For both phases, *alpha*, the penalizing factor determining the degree of the feature selection for the Lasso regression is fitted using five fold cross-validation.

In line with Ris and Sjöberg (2021) and Thorsrud (2016) the 80 topics are reduced by manually selecting and omitting the topics deemed to be unimportant. For example topics 63, 69 and 75 in Figure 3.1 are omitted in the manual feature selection step. This introduce a subjective element in the process of feature selection, and some important topics may be left out. But, based on the result in Ris and Sjöberg (2021) where this approach led to improved predictive results, the potential gain is deemed larger than the risk of leaving topics out. This procedure result in a input data set consisting of 35 topics going into the Lasso regression algorithm. The 35 topics are presented in Table C.1 found in the appendix. The reduced input data set is then standardized, this procedure removes the mean of the features and scale them to unit variance.

3.6 Predictive Modelling

During the validation phase the different input data representations described in Section 3.4 are used to model the economic variables, and the results compared. The training data set is partitioned in two subsets: a sub training set and a validation set acting as pseudo unseen data. To select between the different input data representations the R^2 score on the validation data are compared, between representations and to a baseline. R^2 score is in it self a comparison of a regression to a baseline of just fitting an intercept or a constant. Another common used baseline is an auto-regressive model of order one, AR(1), this model use the lagged response variable at time *t-1* to predict the variable at time *t*, as described in equation (3.2). See for example Larsen and Thorsrud (2019) where AR(1) models are used as baselines. In the testing phase the best models are compared to an AR(1) model for the economic variable.

$$EV_t = \alpha + \beta EV_{t-1} + \epsilon_t \tag{3.2}$$

The economic variables are modelled using two slightly different regressions, shown in equations (3.3), and (3.4). The economic variables are regressed on the lagged news topics (NT), to avoid look ahead bias, and $\tilde{K}$ are the reduced number of news topics in the input data set. The economic variables are modelled in level, and after taking first differences. These two representations are further extended by also including the lagged economic variable in the model in equation (3.4). The

representation using first differences is omitted from further exploration based on the validation phase.

$$EV_t = \alpha + \sum_{n=1}^{\tilde{K}} \beta_n NT_{n,t-1} + \epsilon_t \tag{3.3}$$

$$EV_t = \alpha + \sum_{n=1}^{\tilde{K}} \beta_n NT_{n,t-1} + \beta_{EV} EV_{t-1} + \epsilon_t \tag{3.4}$$

The reason for testing and comparing two different regressions for the input data representations is to investigate whether the newspaper data can be used on a stand alone basis as in Ris and Sjöberg (2021), and if the newspaper data adds to a AR(1) model as in Larsen and Thorsrud (2019). The results from using the model in equation (3.3) will help answer the first question and the results from using the model in equation (3.4) the second.

In the testing phase the monthly distribution data set and the three most promising alternative input data representations from the validation phase are tested on the test data set. These input data representations are described in Table 3.1. Again, the predictive modelling is conducted by Lasso regression utilizing five-fold cross-validation to determine *alpha*, but now training the model on the full training data set and predicting on unseen data.

3.7 Results

The results presented below are the out of sample R^2 scores obtained after using five-fold cross-validation to determine the *alpha* value for the Lasso regression. The obtained configuration is then fed unseen data points to predict the respective target values. The in-sample R^2, on the training data is usually high or at least above zero for most input-output representations. This is true when feeding the model only newspaper data as well as when including the lagged economic variable.

In table 3.3 the out of sample R^2 scores from AR(1) modelling of the economic variables are presented. Relatively low scores can be seen for *Expected Inflation*, *Unemployment* and *CPIF-inflation* when modelled using the normal data set.

Economic Variable	R^2
Expected Inflation	0.04
BCI	0.74
CCI	0.82
Consumption	0.66
Unemployment	0.20
Inflation (CPI)	0.73
Inflation (CPIF)	0.25

Table 3.3: Out of sample R^2 scores for target variables modeled as AR(1).

When only using newspaper data as input data in the predictive model is not showing any promising results. The R^2 score is negative for almost all combinations of input data representations and economic variables. There is however one exception, *CPIF-inflation* modelled using the moving average input data representation. This combination yield a score of 0.38, well above the baseline of 0.25 obtained from the AR(1) model. The result is also well above 0, the baseline obtained from fitting an intercept only. The topics included in the regression by the Lasso algorithm are presented in table C.3. None of the other combinations manages to beat the baseline of fitting an intercept.

When allowing for the model to also include the lagged economic variable as described in equation (3.4) the results improve. These results can be seen in table 3.4. Now there are some combinations beating the baseline score for the AR(1) model, for example *expected inflation* modelled using squared sum representation, *unemployment* modelled using the monthly distribution representation and *CPIF-inflation* using the daily document representation.

Dataset	Squared Sum	Daily Doc.	Moving Avg.	Monthly Dist.
Expected Inflation	**0.31**	-	0.04	**0.14**
BCI	0.71	0.70	0.74	**0.77**
CCI	0.80	**0.83**	0.82	0.82
Consumption	0.48	**0.71**	0.38	0.49
Unemployment	**0.21**	-	-	**0.31**
Inflation (CPI)	**0.75**	0.41	0.73	0.72
Inflation (CPIF)	**0.30**	**0.39**	0.19	-

Table 3.4: Out of sample R^2 scores for target variables modeled as using newspaper data and the lagged variable. A dash (-) marks a negative score, and bold font marks scores above baseline.

Based on the results in Table 3.4 there is not a clear best data representation of the four separate representation compared in the test phase. This is also in line with what is seen in the validation phase, where some representation seem to work better for some economic variables. For expected inflation the squared sum representation seem to add the most information while the daily document representation actually results in a negative score. For BCI and CCI no representation seem to notably improve the results, except for the monthly distribution data set for BCI.

For consumption the daily document representation is the clear best representation. For Unemployment the monthly distribution outperform the other representations, but squared sum leads to a positive result while the others are negative. Square sum performs well for both inflation measures, but is beaten by the daily document representation for CPIF-inflation.

3.8 Discussion

When not allowing the regression to include the lagged variable of interest the results are showing that the current representation of the newspaper data does not add any out of sample predictive power. This is the case for almost all input-output combinations, except for *CPIF-inflation* modelled using the moving average input data representation. But, when allowing this same combination to also include the lagged economic variable the results deters and falls below the R^2 obtained by just using the lagged variable. This is not a nice property, as the result is expected to move in opposite direction when adding a variable actually adding information as can be seen in Table 3.3.

In Ris and Sjöberg (2021) the predictive modelling is carried out only using input data in form of newspaper data representations, with more promising results compared to the results obtained for modelling the economical variables in this work using only text data. The stock market is known for being difficult to predict, but the objective formulation in Ris and Sjöberg (2021) being a binary classification problem and not a regression problem could be a reason impacting the results. When a model prediction for example "overshoots" in a binary classification the model is still right, but when a prediction "overshoots" in a regression this has a negative effect on the result. Another clear distinction between this study and Ris and Sjöberg (2021) is the availability of data, the stock market's closing prices is updated every trading day resulting in many more data points, compared to modelling economic variables with a monthly update frequency. The 120 data points from the ten years of newspaper data used in this work could be a too small sample for the model to learn from. As a 80 % - 20 % train - test split is used this result in only 96 data points to train on and 24 data points to test on.

When allowing the model to train on the lagged economic variable on top of the newspaper data set representation the results are improved, and in some cases not only the intercept baseline is beaten but also the AR(1) baseline. However, the results are not completely convincing using this setup either, for example as there is no consensus regarding the best data representation. But, the results that actually are promising are showing how the addition of the newspaper data set are increasing the R^2 score by a noteworthy amount compared to only using the AR(1) model. This is the case for the variables with lower initial scores: expected inflation, unemployment and CIPF-inflation. There are also minor improvements for other combinations of data representation and economic variable. This result is hinting that Swedish newspaper data modelled as in this work have predictive power for at least some Swedish economical variables. This is in line with (Larsen and Thorsrud,

2019), showing how news topics constructed from Norwegian news are improving AR(1) forecasts for Norwegian economical variables.

While there are difficulties distinguishing a best fit data representation for the newspaper data to combine with a lagged economic variable. At least one could potentially omit the moving average representation, as none of the results obtained when this data representation is combined with the lagged variable were above baseline. However, this representation was the only representation that managed to score an above baseline result on stand alone basis, and could on that ground be explored further, but perhaps not in combination with the lagged variable of interest.

For the business confidence index the newspaper data seems not to add any vital information when comparing to the AR(1) representation. This is a result similar to Larsen and Thorsrud (2019), where very few variables improved the predictions of BCI. This is also the case for the consumer confidence index, originating from the same source as the BCI variable which is seemingly challenging to predict using news data. Also, the predictions of the consumption variable and CPI-inflation are not improved by much using any data representation. This is contrary to Larsen and Thorsrud (2019), where consumption and output were the two variables showed to be connected to most news topics. When looking at the plotted target variables in 2.1, one observation is that the BCI, CCI and consumption time series all are smoother than the rest of the variables, presumably due to being constructed differently.

4

Conclusion

In this study newspaper data from Dagens Industri, Sweden's largest business newspaper is represented using topic modelling and sentiment analysis originating from the field of NLP. Different representations by combination of topics and sentiment scores are constructed and their usefulness are compared for modelling various economic variables for the Swedish economy. The newspaper data on a stand-alone basis have not been shown to result in promising predictions of these macroeconomic variables. But, when also including the lagged variable of interest in the regression there are cases where the newspaper data increase the predictive power compared to only using the lagged variable. The three economic variables where the inclusion of newspaper data are shown to be help explain most variability are: *expected inflation*, *unemployment* and *CIPF-inflation*.

The objective of this study is to explore the usage of Swedish newspaper data represented by LDA and sentiment analysis for prediction of Swedish economic variables. This work has at least to some extent reinforced earlier findings that news topics improve the predictive power of an AR(1) model for some economic variables, but now also for the Swedish economy. However, the study being of an explorative nature the results are indicative and more research is necessary to show this in a more rigorous setting. For example as the topic modelling in this study is conducted for the full data set, introducing a potential look ahead bias in the results as the topics are constructed using data from the test period as well as the train period.

In this work the ten years of newspaper data are boiled down to 120 data points due to the target data being on monthly basis. Future research of the usage of newspaper data for predictions in the field of economics could explore variables with a higher update frequency, as this would allow for a larger data set with more data for the model to train on. Another idea for increasing the data points is to increase the length of the newspaper sample, conducting similar research as in this work but utilizing a 30 year period would result in three times more data to train on. This would of course also increase run times and complicate data handling and processing in the data preparation phase, but could be a viable solution. More data could also allow for the usage of more sophisticated machine learning algorithms, which could show to be more effective.

In this study the full length data set is included in the topic modelling, this means text data from both the train and test period. Future studies could conduct the topic modelling by exclusively letting the model train the data from the training period. Then, using this topic model to infer topics for the test period and examine how this affect the results. Such approach would completely erase the potential look ahead bias present in this work. An alternative, but computationally costly approach, would be to utilize continuous updating of the topic model and retraining of the predictive model.

This work explore only the usage of topic modelling and sentiment analysis for text representations. There are many other ways to represent text data that could show to be fruitful for macroeconomic predictions. Exploring such other representations could be an objective of future studies, as this study is indicative of the usefulness of newspaper data but have not found a clear best representation, not even within the ones compared.

As there seems to be an increase in predictive power, combining Swedish newspaper data and a lagged economic variable, an idea for future works is to extend this method in other fields. For example could stock market predictions conducted using newspaper data on a stand alone basis, potentially benefit from combining the newspaper data with other variables of interest and possibly obtain even better results. Another idea for future works within economics could be to do the opposite, to augment existing data sets created for prediction of economic variables with newspaper data.

References

J. Blad and K. Svensson. Exploring nmf and lda topic models of swedish news articles. Master's thesis, Uppsala University, Uppsala, 12 2020.

D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993, 01 2013.

D. Bzdok, N. Altman, and M. Krzywinski. Statistics versus machine learning. *Nature methods*, 15(4):233–234, 4 2018.

D. Chicco, M. Warrens, and G. Jurman. The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation. *PeerJ Computer Science*, 7:e623, 7 2021.

Dagens Industri. Dagens industri: Om oss. URL https://www.di.se/nyheter/om-oss/, 2020. Accessed: 2021-06-19.

M. Gerlach, H. Shi, and L. A. N. Amaral. A universal information theoretic approach to the identification of stopwords. *Nature Machine Intelligence*, 1:606–612, 12 2019.

Konjunkturinstitutet. Statistikdatabasen. URL https://www.konj.se/statistik-och-data/statistikdatabasen.html, 2021. Accessed: 2021-06-27.

S. L. Kristiansen. Lemmy. URL https://pypi.org/project/lemmy/, 2019. Accessed: 2021-07-25.

E. Kumar. *Natural Language Processing*. I.K. International Publishing House Pvt. Limited, 2011. ISBN 9789380578774.

V. H. Larsen and L. A. Thorsrud. Asset returns, news topics, and media effects. Working Papers No 5/2017, Centre for Applied Macro- and Petroleum economics, BI Norwegian Business School, 10 2017. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=305795.

V. H. Larsen and L. A. Thorsrud. The value of news for economic developments. *Journal of Econometrics*, 210:203–218, 5 2019.

V. H. Larsen, L. A. Thorsrud, and J. Zhulanova. News-driven inflation expectations and information rigidities. *Journal of Monetary Economics*, 117, 3 2020.

T. Loughran and B. Mcdonald. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66:35 – 65, 2 2011.

A. K. McCallum. Mallet: A machine learning for language toolkit. URL http://mallet.cs.umass.edu, 2002. Accessed: 2021-06-22.

K. P. Murphy. *Machine learning: a probabilistic perspective*. MIT Press, 2013. ISBN 9780262018029. Available Online. URL https://mitpress.mit.edu/books/machine-learning-1.

OECD. Business confidence index (bci) for sweden. URL https://data.oecd.org/leadind/business-confidence-index-bci.htm, 2021a. Accessed: 2021-07-10, Select Sweden.

OECD. Consumer confidence index (cci) for sweden. URL https://data.oecd.org/leadind/business-confidence-index-cci.htm, 2021b. Accessed: 2021-07-10, Select Sweden.

OECD. Harmonized unemployment rate: Total: All persons for sweden. URL https://fred.stlouisfed.org/series/LRHUTTTTSEM156S, 2021c. Accessed: 2021-07-10, Retrieved from FRED.

Retriever. Retriever mediearkivet. URL https://www.retriever.se/tag/mediearkivet/, 2021. Accessed: 2021-02-21.

E. Ris and A. Sjöberg. Index prediction on the swedish stock market using natural language processing methods on swedish news. Master's thesis, Lund University, Lund, 6 2021.

A. Shapiro, M. Sudhof, and D. Wilson. Measuring news sentiment. *Journal of Econometrics*, 11 2020.

Svenskt Näringsliv. Om oss. URL https://www.ekonomifakta.se/mer/Om-Ekonomifakta/, 2015. Accessed: 2021-07-12.

Take Lab. Spacy-udpipe. URL https://pypi.org/project/spacy-udpipe/, 2020. Accessed: 2021-03-10.

L. A. Thorsrud. Nowcasting using news topics. big data versus big bank. *SSRN Electronic Journal*, 1 2016. doi: 10.2139/ssrn.2901450.

L. A. Thorsrud. Words are the new numbers: A newsy coincident index of the business cycle. *Journal of Business and Economic Statistics*, 38:1–35, 8 2018.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58:267–288, 1 1996.

Appendix A

Natural Language Processing

NLP is a research field with roots in linguistics as well as in data science. The overall goal of NLP is to make sense of textual language data by representing it for automatized processing by computers, (Kumar, 2011).

A.1 Text Data Processing

A common and vital approach before any advanced NLP techniques are applied to a text data set is to "clean" or pre-process the text data. There are many techniques available to apply, and can be said to be more art than science. Some of the more common approaches are listed below.

To clean the text, some characters such as numbers, punctuation, special characters (for example: #,$ or @) are removed. This step could also include removal of words such as URL-links and stop words (i.e. very common words carrying little information such as: and, so, in). Often words that occur very frequently or very rarely are omitted from the data set in a final stage, this however, is a debated subject and with much current research (Gerlach et al., 2019).

In order to having only one representation of each word the text is transformed to lower case. And, to further increase this effect methods such as stemming: transforming a word into its word stem, or lemmatization: transforming a word into its lemma are often used. This yields a data set where each word is mapped to the same lemma or stem regardless of tense or conjugation.

A.2 Text Data Representation

In order to process a text data set further and potentially extract more information, *n-grams* can be introduced to the text. N-grams of level two, two-grams or bigrams, are constructed by letting two words following each other make up a new token. For

example the word sequence "semi automatic machine gun" will result in the following two-grams: "semi automatic", "automatic machine" "machine gun". Where the first and the latter two-grams might add more information compared to the individual words alone. Adding n-grams will make a bag of words (BoW) based model such as LDA able to interpret connected words as one entity. This is method adds more information to languages such as English where most words are separated by a space, but is often also used for other languages.

In order to represent different text documents in a corpus a common approach is a BoW representation. To create a BoW representation of a corpus, all documents are decomposed into their respective words, creating a set of all possible words in the corpus. Each document is then represented by one-hot encoding, depicting which of all possible words the document contains. If, for example a corpus consists of two documents: doc_1: "I like football" and doc_2: "You like coffee", then all of the possible words in the corpus are: I, like, football, you, coffee and the two document are represented in a BoW format as shown in A.1.

Document	I	you	like	football	coffee
doc_1	1	0	1	1	0
doc_2	0	1	1	0	1

Table A.1: Bag of words representation of two example documents.

A.3 Topic Modelling

In order to create a interpretable representation of a text corpus topic modelling can be used, (Ris and Sjöberg, 2021). The purpose of topic modelling is to decompose a text corpus into the a predefined number of topics, K. There are different viable algorithms to use in order to achieve this but one of the more frequent used and accessible algorithms is LDA. (Larsen and Thorsrud, 2017) Another common topic modelling approach is Non-negative Matrix Factorization (NMF), both methods are compared for Swedish newspaper data in Blad and Svensson (2020).

LDA, first presented in Blei et al. (2013), is a latent variable topic model, where the latent (non-observable or not yet found) variables of the model are what humans would recognize as topics, and the observable variables are the structure of words and documents. Each document is assumed by the model to be constructed by a generative process. A document is presumed to be built up by an on beforehand specified number of topics, which are in turn constructed by a mix of words. Each word in each document is assumed to belong to one of these K topics. The purpose of the LDA model is to extract the topics, from a corpus consisting of documents and words. This is done by inferring the conditional distribution of the non-observed variables, given the observed variables, utilizing their joint probability distribution. The distribution of the topics within documents and distribution of words within topics are both thought to be Dirichlet distributions. What the LDA model is set out to do is to approximate the conditional probability of the topics given how the documents are structured. There are different ways to accomplish this, one could

take a sampling based approach as the Gibbs sampling in the implementation of the LDA algorithm used in this work, or a optimization approach such as Online Variational Bayes. In Ris and Sjöberg (2021) a technical explanation of the math behind the LDA-algorithm is presented.

LDA builds upon the idea that a text corpus is constructed by a set of M documents and each document by a set of words. The vocabulary V is all of the words in the corpus. Each topic is described as a distribution over the the words in the vocabulary, assigning a probability for each word to belong to that specific topic. It is important to note that a LDA model does not account for the syntax or order of words in documents and treat all documents as a set of unordered words. The input data for the LDA model is therefore a bag of words representation of the documents in the corpus, for a newspaper data set usually the articles. The output of the model is the word distribution for the topics given by a a $K \times V$ matrix, and a topic distribution for the documents given by a $M \times K$ matrix. Often, topics generated by a topic model are presented as the most important words for each topic, instead of the full word distribution.

For an illustrative example, remember the two documents: doc_1: "I like football" and doc_2: "You like coffee". Let these two documents construct a corpus, M is also in this case 2. If an LDA model, specified to find 4 topics ($K = 4$) for this corpus these topics could be named and described by a distribution over all unique words in the vocabulary V, seen in Table A.2. The figurative topic distribution for the documents can be seen in Table A.3.

Topic	I	you	like	football	coffee
"Pronouns"	0.31	0.39	0.1	0.1	0.1
"Liking"	0.03	0.07	0.8	0.05	0.05
"Coffee"	0.05	0.05	0.05	0.02	0.83
"Football"	0.01	0.02	0.02	0.9	0.05

Table A.2: Example of figurative topics generated by a LDA model described by their word distribution.

Document	"Pronouns"	"Liking"	"Coffee"	"Football"
Doc_1	0.3	0.3	0.1	0.3
Doc_1	0.3	0.3	0.3	0.1

Table A.3: Example of a figurative topic distribution matrix generated by a LDA model.

One approach to getting a better understanding of how the LDA algorithm works is to imagine how the documents in a corpus could have been generated, (Larsen and Thorsrud, 2019). Assuming all topics are known, then the documents will be generated by first picking the overall theme of the document as a random distribution over all topics. Then, for each word the article should contain randomly pick a topic from the document's topic distribution, and from that topic's word distribution randomly choose a word. Iterating this procedure should result in a corpus of documents constructed by words from many different topics. However, some topic is more important than the others for all articles, and can be said to describe this

document. When the LDA topic model is identifying and extracting topics from a corpus the method is the inverse of this generative process.

A.4 Sentiment Analysis

Sentiment analysis in a NLP context aims to find the tone of a text. (Shapiro et al., 2020) This can be defined in different ways using many dimensions. A simple and straightforward approach is to deem if a text is positive or negative, this can be further extended by introducing a scale dimension, i.e. "how positive or negative is this text?".

There are two main approaches to sentiment analysis, one being a dictionary based and one being a machine learning based. A dictionary based approach is using a look up table consisting of words or phrases on beforehand classified as corresponding to a certain sentiment measure. A machine learning based approach utilize classification algorithms, and thus, needs to be provided labeled examples of texts to train on and learn from before it can be used to classify new documents.

Appendix B

Machine Learning

B.1 Overview

Machine learning is traditionally divided into supervised learning, unsupervised learning and reinforcement learning. The latter of these will not be discussed further, as no such methods are explored in this work. Machine learning is traditionally more inclined towards making predictions compared to traditional statistics where more focused lies on establishing relation. (Bzdok et al., 2018) The lingo or terminology used in machine learning and statistics is slightly different, which is something that can create confusion. (Murphy, 2013) For example, what is known as "explanatory variables" or "independent variables" in statistics and econometrics is called "features" in machine learning literature.

Supervised learning is a subbranch of machine learning used when the data set at hand have predefined labels. An example of a simple supervised machine learning algorithm is *linear regression* where you have both input data in form of x-values and output data in form of y-values, for example mapping *height* to *age*. In supervised machine learning the models train on labeled data and are evaluated by comparing the model output to the true labels, simplifying model evaluation.

In unsupervised learning there are no correct labels for the data set, and the purpose of the algorithms used are to find relations or groups in the data. Examples are *LDA* used to find K topics in a large corpus of texts and *K-means* which aims to find K clusters in a data set. For both algorithms K is specified on beforehand without any information on the actual topics or clusters. This makes model evaluation for unsupervised learning models a more difficult task compared to supervised learning.

B.2 Feature Engineering

In statistical modelling and machine learning a common practise is to scale the input features. This is performed to increase performance and simplify learning. This as

the range and magnitude of the input features could have a high impact on the cost function to be optimized, for example if this function is derived from minimizing a euclidean distance.

There are two common ways to scale input data: *standardization* and *normalization*. When an input feature is normalized it is scaled and shifted to fall in a predefined range, often [-1,1] or [0,1], and when an input feature is standardized the data are converted to a zero mean distribution with unit variance, but this distribution does not have to be Gaussian.

B.3 Predictive Method and Feature Selection

Lasso regression is a regression algorithm simultaneously performing feature selection and regularization, (Tibshirani, 1996). The algorithm is often used with the purpose of simplifying a large feature space in order to increase interpretability. By utilizing the L1 norm for regularization the algorithm is able to completely omit unimportant features by setting their coefficients to zero. The objective function of Lasso regression can be seen in equation (B.1).

$$\min_{w} \frac{1}{2n} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}||_2^2 + \alpha ||\boldsymbol{w}||_1 \tag{B.1}$$

Cross-validation is an approach taken in order to reduce overfitting and improve the overall fit of the model, often also used to fit hyperparameters of a machine learning model. In traditional K-fold cross-validation the data set is partitioned in K folds or sets, and the model is trained on *K-1* of these and evaluated on the last remaining set. This is repeated K times, evaluating one time for each set and thus, resulting in K evaluation metrics. The overall evaluation metric of the cross-validation sequence is then the average of the K evaluation scores. Often K is set to five, resulting in a sequence described in Table B.1.

5-fold cross-validation					
Iteration	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
1	Train	Train	Train	Train	Validate
2	Train	Train	Train	Validate	Train
3	Train	Train	Validate	Train	Train
4	Train	Validate	Train	Train	Train
5	Validate	Train	Train	Train	Train

Table B.1: Description of the process for five fold cross-validation.

B.4 Result Interpretation

R^2 or the coefficient of determination is a measure of the amount of the variability of the response variable that can be predicted by the explanatory variables. The definition is presented in equation (B.2). A baseline model always predicting the mean of the response variable would yield a R^2 value of 0. As suggested by (Chicco et al., 2021) R^2 is an informative and truthful measure for regression analysis evaluation and to be preferred over other common evaluation metrics such as mean average percentage error (MAPE) or root mean square error (RMSE).

$$R^2 = 1 - \frac{SS_{tot}}{SS_{res}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{B.2}$$

When interpreting results of any test a baseline is required to be able to compare the obtained result and deem if the test is successful. R^2 as mentioned in the former paragraph compare the obtained model with only fitting an intercept. A constant or an intercept could function a simple and interpretable baseline for a regression analysis.

Appendix C

Topics in Swedish

C.1 Topics described by Wordclouds

Figure C.1: Topic 0 - topic 19 presented as Wordclouds in Swedish.

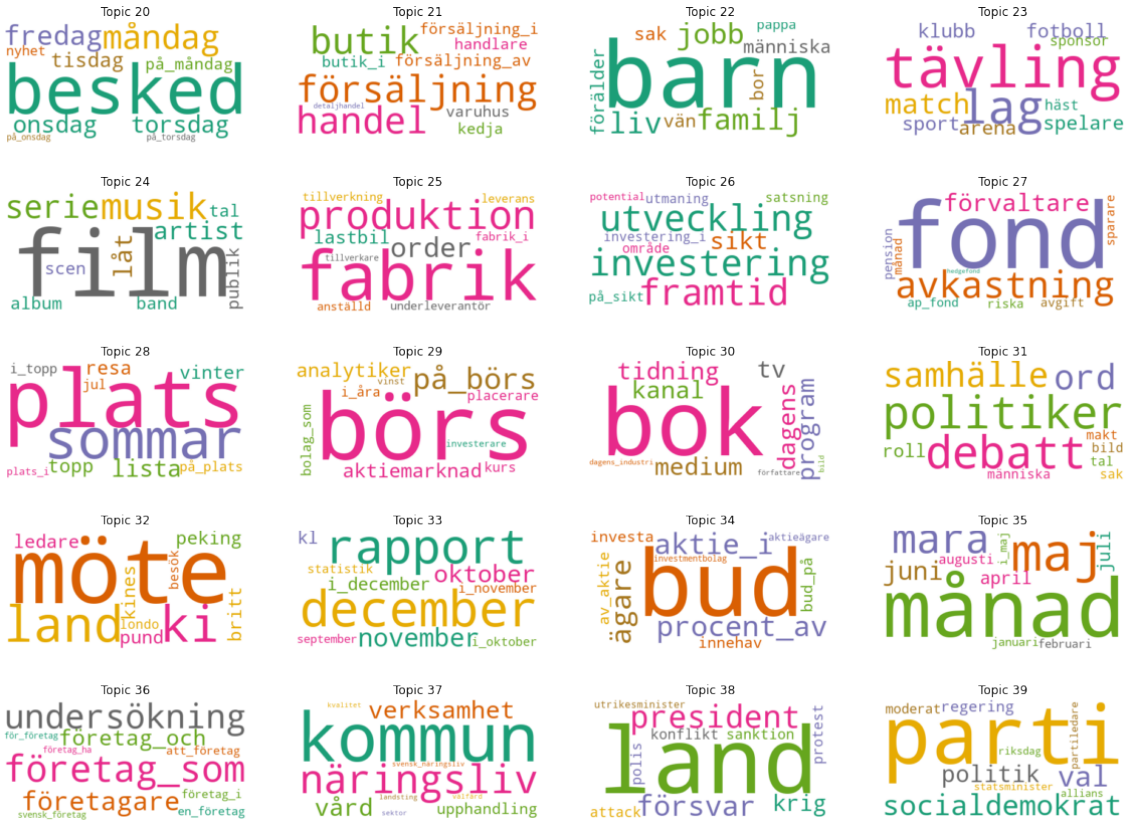

Figure C.2: Topic 20 - topic 39 presented as Wordclouds in Swedish.

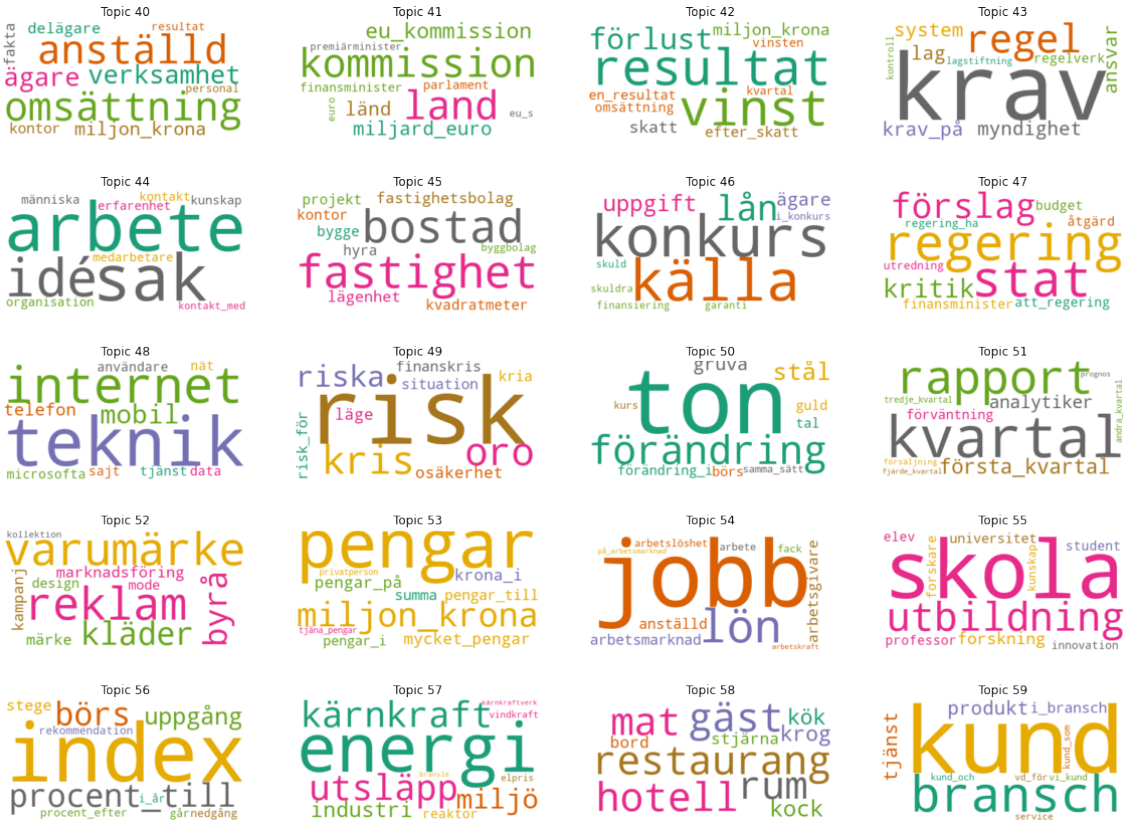

Figure C.3: Topic 40 - topic 59 presented as Wordclouds in Swedish.

Figure C.4: Topic 60 - topic 79 presented as Wordclouds in Swedish.

C.2 Interesting Topics

Topic	Important words
1	utdelning, kapital, värdering, aktieägare, tal, egen_kapital, värde
4	affär, köp, försäljning, köpare, köp_av, försäljning_av, säljare
5	utdelning, tal, handel, prognos, räkenskapsår, antal, vinst
6	skatt, inkomst, skatteverk, skatt_på, avdrag, pensionär, avgift
13	tillväxt, verksamhet, lönsamhet, omsättning, koncern, förvärv, vd_n
15	länd, land, export, region, i_land, andra_länd, länd_som
16	ränta, riksbank, centralbank, inflation, kronan, valuta, styrränta
17	dollar, olja, oljepris, miljard_dolla, dollar_per, oljebolag, produktion
21	försäljning, butik, handel, försäljning_av, försäljning_i, kedja, butik_i
25	fabrik, produktion, order, lastbil, tillverkning, fabrik_i, underleverantör
26	investering, utveckling, framtid, sikt, investering_i, utmaning, satsning
27	fond, avkastning, förvaltare, avgift, ap_fond, sparare, pension
37	kommun, näringsliv, verksamhet, vård, upphandling, svensk_näringsliv
38	land, president, försvar, krig, konflikt, sanktion, polis
39	parti, socialdemokrat, val, politik, regering, moderat, statsminister
40	omsättning, anställd, ägare, verksamhet, miljon_krona, delägare, kontor
41	kommission, land, eu_kommission, miljard_euro, länd, finansminister
42	resultat, vinst, förlust, miljon_krona, skatt, efter_skatt, en_resultat
45	fastighet, bostad, fastighetsbolag, lägenhet, kvadratmeter, projekt
46	konkurs, källa, lån, uppgift, ägare, i_konkurs, finansiering
47	regering, stat, förslag, kritik, att_regering, finansminister, budget
48	teknik, internet, mobil, telefon, tjänst, nät, sajt
49	risk, kris, oro, riska, osäkerhet, läge, finanskris
50	ton, förändring, stål, gruva, förändring_i, tal, guld
54	jobb, lön, arbetsmarknad, arbetsgivare, anställd, arbetslöshet, fack
55	skola, utbildning, forskning, universitet, professor, student, elev
57	energi, kärnkraft, utsläpp, miljö, industri, reaktor, vindkraft
60	läkemedel, apotek, patient, forskning, hälsa, patent, test
64	president, val, republikan, kandidat, kongress, delstat, demokrat
68	analys, nyheter, börs, handel, option, nasdaq, antal
71	bank, lån, storbank, att_bank, bank_och, utlåning, kapital
73	pris, hushåll, pris_på, bostadsmarknad, bostad, att_pris, pris_för
77	investerare, riskkapitalbolag, kapital, nyemission, bolag_som, notering
78	ekonomi, tillväxt, prognos, återhämtning, nästa_åra, konjunktur, tecken
79	hq, domstol, utredning, information, granskning, brott, revisor

Table C.1: The 35 topics in the reduced data set constructed by omitting uninformative topics from the original 80 topics. The topics are described by their six or seven most important words, given in Swedish.

Topic	Important words
46	konkurs, källa, lån, uppgift, ägare, i_konkurs, finansiering
71	bank, lån, storbank, att_bank, bank_och, utlåning, kapital
78	ekonomi, tillväxt, prognos, återhämtning, nästa_åra, konjunktur

Table C.2: Topics 46, 71 and 78 described by their seven most important words, given in Swedish.

Topic	Important words
1	utdelning, kapital, värdering, aktieägare, tal, egen_kapital, värde
4	affär, köp, försäljning, köpare, köp_av, försäljning_av, säljare
5	utdelning, tal, handel, prognos, räkenskapsår, antal, vinst
6	skatt, inkomst, skatteverk, skatt_på, avdrag, pensionär, avgift
17	dollar, olja, oljepris, miljard_dolla, dollar_per, oljebolag, produktion
21	försäljning, butik, handel, försäljning_av, försäljning_i, kedja, butik_i
25	fabrik, produktion, order, lastbil, tillverkning, fabrik_i
26	investering, utveckling, framtid, sikt, investering_i, utmaning
39	parti, socialdemokrat, val, politik, regering, moderat, statsminister
40	omsättning, anställd, ägare, verksamhet, miljon_krona, delägare
41	kommission, land, eu_kommission, miljard_euro, länd, finansminister
45	fastighet, bostad, fastighetsbolag, lägenhet, kvadratmeter, projekt
47	regering, stat, förslag, kritik, att_regering, finansminister, budget
48	teknik, internet, mobil, telefon, tjänst, nät, sajt
50	ton, förändring, stål, gruva, förändring_i, tal, guld
54	jobb, lön, arbetsmarknad, arbetsgivare, anställd, arbetslöshet, fack
57	energi, kärnkraft, utsläpp, miljö, industri, reaktor, vindkraft
60	läkemedel, apotek, patient, forskning, hälsa, patent, test
64	president, val, republikan, kandidat, kongress, delstat, demokrat
68	analys, nyheter, börs, handel, option, nasdaq, antal
71	bank, lån, storbank, att_bank, bank_och, utlåning, kapital
73	pris, hushåll, pris_på, bostadsmarknad, bostad, att_pris, pris_för
77	investerare, riskkapitalbolag, kapital, nyemission, bolag_som, notering
78	ekonomi, tillväxt, prognos, återhämtning, nästa_åra, konjunktur

Table C.3: Topics included in the regression by the Lasso algorithm modelling CPIF-inflation using the moving average input data representation. The topics are described using six or seven important words, given in Swedish.